Private Persons and Minimal Persons*

Elijah Millgram

Department of Philosophy
University of Utah
215 Central Campus Drive, CTIHB 4th floor
Salt Lake City UT 84112

lije@philosophy.utah.edu

April 4, 2012

It's a commonplace that privacy can now be abridged and abdicated in ways that weren't routinely possible until very recently. I want here to draw attention to an alternative configuration of the mind that these techniques make available, which I will call the *minimal person*.

My explication of minimal personhood is going to take the long way around. I will have to explain what the ethical and political concept of privacy has to do with the older and very different philosophers' notion of logical privacy: this part of the discussion will connect the recent debates over extended cognition and first-person authority to one another. To get into a position where I can do that, I will have to explain how personhood and the laws of logic are also related topics. And to do *that*, I will start out with an exercise in what Paul Grice and, following him, Michael Bratman have called 'creature construction.'[1]

1

Creature-construction arguments proceed by describing a series of progressively more ambitious organisms or robots, each of which handles a perfor-

*I'm grateful to Chrisoula Andreou and Leslie Francis for comments on an earlier draft, to Meg Bowman, Michael Gill, Matt Haber, David Humphries, Jenann Ismael, Kimberly Johnston, Jan Murphy, Shaun Nichols and Dan Russell for helpful conversation, and to the Center for the Philosophy of Freedom at the University of Arizona for work space.

[1]Grice, 1975; Bratman, 2006, *passim* and esp. 49f; Millgram, 1997, ch. 5, is a creature-construction argument on a topic related to the treatment in the present paper.

mance shortfall identified in its predecessor; the immediate point of these descriptions is to isolate the features that support the incremental improvements in performance, and thereby to justify incorporating those features into the design of an agent faced with a particular range of challenges. After we have the upcoming creature-construction argument in place, I will return to the question of how we are supposed to understand their upshots.

I am going to opt for imaginary robots over imaginary organisms, and the primary dimension along which I am going to arrange my robots is how much the robotics design team knows about the environments in which they are intended to operate.[2] Accordingly, at the beginning of the series, we can place the Mark I, which is constructed for a thoroughly understood and fully predictable environment: perhaps an assembly line on which regularly spaced, identically positioned widgets undergo a mechanically identical modification. For this task, the sequence of motions of the robot's arm and gripper can be hard-coded into silicon, and consequently the Mark I needs scarcely anything in the way of sensory inputs; it does not, for instance, need to see the oncoming widgets. And it needs scarcely anything in the way of a representation of its environment. ('Scarcely anything': the designers might want to build in a series of checkpoints, e.g., to verify that the gripper did close down on the part.) Because robots are thought of as autonomous (at least to some extent), this is a trivial and limiting case of an industrial robot.

But of course the Mark I is usable only in an extremely controlled environment. If a human being further up the line is putting the half-assembled widgets onto the conveyor belt any old which way, the Mark I must be junked, and the factory will upgrade to the Mark II: a substantially more sophisticated device equipped with a camera, capable of determining the positioning of a widget, and capable of computing the movements of its manipulator. The Mark II will construct and update representations of the positions of the oncoming widgets, of its own effector, and perhaps other elements of its environment. However, those quite limited representations will not yet amount to a map—say, of the factory floor. The important step that has been taken here is that of delegating to the device part of the control that, in the Mark I, could be exercised directly by the robot's designer: only a small part, however, since in this case the designer is able to anticipate a great deal of what the robot is and is not to do. For instance, it is normal in robots of this type to manually specify a volume outside of

[2]Thus this particular creature-construction argument is very much in the spirit of Ismael, 2007.

UU IR Author Manuscript

which the robot's arm may not move.

The design team might decide that it wants a robot, the Mark III, that can be repositioned at different points in an assembly process. The Mark III, unlike its predecessors, is able to move around the factory floor, and so now a map is necessary. Perhaps the team can have the robot load a floor plan they provide during its initialization procedure, but because equipment is expected to move, the Mark III will need to update its map; moreover, because avoiding industrial accidents is a priority, it is a good idea to allow for updating in real time, on the basis of sensor inputs. (If the robot bumps into something that its map says is not there, it should be able to overwrite the map entry.) And because the programmer does not know where the robot will be at any particular time, he will have to delegate to the robot still more of the task of determining what it is going to do: now the robot must compute how it is to execute its tasks, given an initial location, and a map of the factory floor which it is responsible for updating.

There are many such incremental improvements to be walked through, at each of which the design team hands off more control to the robot it is designing. In the interest of brevity, I'm going to skip ahead a few models to a technique which I'm mentioning primarily because of its familiarity to philosophers. In one way of controlling the activities of a robot, let's say it's the Mark N, it derives a series of operations to execute via backward chaining from a specified set of goal states.[3] There are a great many philosophers who seem to think that such means-end derivations are the only form of problem solving through which the solution to an action control problem can be computed.[4] In fact, programming methods and control techniques vary, and here we can see this one to be just another step in such a sequence of designs: the step at which the engineer reserves specification of goals as his own prerogative, but delegates computation of a sequence of operations that will attain those goals to the device. The point of doing it this way is that the robot will be functioning in an environment that the designer understands well enough to assign objectives to his robot, but does not understand well enough to be able to specify, ahead of time, how those objectives are to be achieved.

I want to emphasize that there is nothing special about this stage of the sequence, and just to make sure the point sticks, let's also sketch the Mark N+1, designed for a somewhat more demanding environment, in this case, a robotics competition with many, many rounds. At each round, the

[3] SHRDLU was a famous early softbot of this kind. See Winograd, 1972, pp. 117–123.
[4] Smith, 1987, is a well-known example.

UU IR Author Manuscript

robots are scored by a panel of judges on a number of dimensions—but while the dimensions are labeled ('versatility,' 'intelligence,' etc.), the design team can no longer predict what implementable objectives will guide the robot successfully through the contest. Their solution is to allow the scoring, at each stage of the competition, to modify both their robot's objectives, and the priority ordering among them. Each objective is given its own 'bank account,' and whenever it is successfully achieved, its account is augmented by the panel's score; objectives bid, in a simulated auction, to control the robot's behavior during the next stage of the competition; at every round, some percentage of the objectives mutate. Over the course of the competition, the robot's mix of objectives evolves to match the contours of the judges' implicit standards.[5] Here the design team understands the environment well enough to identify elements of a behavior-guiding feedback loop, but not well enough to be able to assign objectives to the robot ahead of time, and so they delegate to the robot the task of computing its own goals on the basis of values taken by suitable variables in the loop.

Let's skip ahead a number of stages, to robots intended for still more challenging environments (and now we are pushing past the frontiers of contemporary robotics). The designers will have to delegate the task of computing goals or objectives to the robot in even less structured environments than that of the Mark N+1. To make this vivid, let's imagine that, just as NASA's Jet Propulsion Laboratory has been sending Mars Rovers to a nearby planet to see what's out there, the space agency on some other planet is working on a mobile robot that they are going to use to explore the Earth. Let's call it the *Earth Rover*. What should the Earth Rover look like?

Because the point of such a robot is, precisely, to explore its environment, its designers know very little about what it's like where it's going. So the Earth Rover cannot come equipped with a map it loads from an initialization file; instead, it will have to be its own cartographer. It will also have to develop representations of its environment that don't function in a map-like way; for example, when it records that the larger inhabitants of the Earth seem to be bipeds and quadrupeds, that observation is not associated with any particular map coordinates. Let's refer to this class of data as the Earth Rover's theory of its surroundings. The task of map and theory construction is continuous with the activities of earlier robots in our series, but, obviously,

[5]The approach is adapted from the induction algorithm described at Holland *et al.*, 1986, pp. 47–50, 116–134, 146–150; one of the more interesting aspects of their technique is the way in which credit for success percolates back to upstream nodes in the network.

UU IR Author Manuscript

much more demanding; I will presently consider some of the ways in which it is harder.

How will the Earth Rover make decisions? The Earth is (from the point of view of the design team: is likely to be) a patchwork of very varied landscapes, and a well designed Earth Rover ought to do very different things in the different landscapes. For instance, if it lands in the Sahara, it should take soil and rock samples; but if it lands in the rain forest, it should take foliage samples; and if it lands on the Upper West Side, it should take sushi samples. Maybe its designers know about rocks, but they don't know about rain forests or Manhattan sushi bars ahead of time, and because the robot's home planet is so far away from the Earth, the robot cannot be built around the assumption that it will get timely guidance from its mission control when it first encounters them. Apparently, such a robot must be able to set its own objectives, as it develops its map and accompanying theory of where it has landed.[6] But how can the design team equip a robot to make headway on the parallel mapping and planning problems they are envisaging?

Here is an approach that is a likely and promising part of the architecture for such a robot. (I am not suggesting that it will suffice on its own.) The development team already knows that the robot will be constructing a map

[6]Philosophers too often have entrenched reflexes that need to be turned off before philosophical headway can be made, and the reflex most salient in this case is insisting on a high-level objective or end, one which can be set by the programmer and not delegated. ('Exploring' would be a likely proposal, in this case; 'scoring high' is a likely suggestion for the Mark N+1.) But such broad, cover-all-bases objectives do not support backward chaining: in more traditional philosopher's jargon, when your end is something on the order of 'exploring,' you're not yet in a position to find a means to your end; you must first engage in some form of deliberation of ends. In Aristotle's version of the problem, everyone wants a happy, well-lived life—but what *is* a well-lived life? Before proceeding to try to live one, you have to do something else: perhaps specify, in a much more concrete way, what living well (and likewise: exploring, or performing well by the judges' lights) comes to—and perhaps also what it comes to in *these* circumstances. What this entails is that such ends are formally very different from the goals that do anchor means-end reasons. As a class, they are not well-understood: we don't know how to build robots that act on such ends, and we aren't able to say clearly how we use them ourselves. (For an overview of the specificationism literature, see Millgram, 2008.)

In any case, it is a mistake simply to assume that such engineering problems are to be solved by introducing a high-level or a 'final' end as a reference point, as opposed, say, to constructing feedback loops. E.g., if we define a global variable whose value depends on how much has just been accomplished in the way of additions or changes to a map; if we make the ways in which the robot sets itself mapping tasks sensitive to the value of the global variable; if we do all of that right, we could tell our funding agency that we were building a robotic model of curiosity: and the robot's exploration of its environment would not be, formally, goal- or end-directed. (For some discussion of one such feedback loop in human beings, see Millgram, 2005, ch. 2.)

of the environment it is traversing. They know it will developing a theory to accompany its map. They also know that the robot will maintain data structures that guide its activities: perhaps task schedules or prioritized lists of objectives or variables that play designated roles in feedback loops. To have some shorthand for this, let's say that the robot is maintaining a field of representations. At any point in its mission, the Earth Rover will have available to it an already constructed field of representations.

Now, the design team can be reasonably sure that the Earth Rover's field of representations will contain errors: after all, even the best professional cartographers make mistakes, and who has not found mistakes in his to-do list? Again, the robot will be far, far away on Planet Earth; its mission control will be unable to fix those mistakes, and the robot will have to correct them itself. But those very mistakes can be exploited; in fact, turned into the fuel on which the robot's cognitive processing runs. To mine its mistakes, the design team can give the Earth Rover a set of rules that enforce informational hygiene on its field of representations. Hygiene violations are defined to catch what the design team anticipates are likely to be errors in the evolving field of represenations; for the most part, they trigger one or another procedure meant to identify and rectify the problem. (But hygiene rules need not always target such errors directly: for example, the high-payoff hygiene violations are easier to identify when encoding conventions are uniformly followed, and so the team might determine inconsistent encoding to be itself a hygiene violation; we impose this sort of requirement on ourselves, a trivial example being our insistence on standardized spelling.)

What sort of triggers can the design team anticipate are likely to indicate mistakes in the map under construction? Perhaps they know enough about the surface of the Earth to expect that if the robot has visited the same map coordinates twice, and recorded a different altitude each time, or very dissimilar imaging, something has gone wrong. (Mostly they will be right, but not always: maybe the first time around, the Earth Rover has found its way to the top of an overhanging cliff, and the second time it is at the bottom, under the overhang.) So the robot's informational hygiene will require that each point on the map be tagged with no more than one altitude. Objectives in the task list that are not paired up with plans for accomplishing them might count as hygiene violations. And—again I'm going to provide this example primarily because it will be familiar to philosophers—the robot might have in its theory a pair of representations with the respective forms 'p' and '$\sim p$'; the design team might also designate that as a hygiene violation. They're pretty sure that, when it happens, something has gone wrong, although, from so far away, they won't necessarily know what.

UU IR Author Manuscript

Let's consider that last example a little further. There are evidently many responses the Earth Rover design team might provide to such a violation of representational hygiene. The crudest way of resolving the issue is to delete one of the conflicting representations at random; but the chances of guessing right are no better than fifty-fifty, and in any case, this procedure fails to leverage the hygiene violation to correct less obvious errors in the field of representations. If it is a matter of what the Earth Rover has marked on its map, the robot could revisit the relevant map coordinates and take another look: this will involve adding an objective to the task list, planning a route, and executing the plan. Because such investigations are expensive, however, the design team might see if the error can be diagnosed and fixed without spending valuable surveying time. If memory is cheap enough, the robot can be built to save backups of its representational field, to try to locate the source of the error in its history, and to revise its current informational state on the basis of those diagnostics.

In this approach, computing new objectives can be just a special case of rectifying hygiene violations. For instance, certain kinds of gap in the robot's map might be designated as violations. Sometimes the robot will be permitted to resolve them through interpolation; however, when the data for the surrounding regions are insufficiently smooth, the Earth Rover can be required to assign itself a new objective: that of conducting on-site observations. (When it's a gap in the theory rather than the map, we philosophers will call it induction rather than interpolation, and a very sophisticated version of such a robot may conduct observations elsewhere, meant to support better interpolations or inductions.) Again, when there are too many objectives on its task list to be integrated into an executable plan, that may count as a hygiene violation; one way to resolve it is for the Earth Rover to generate plans that include different objectives from the list, to rank the plans using an appropriate coherence metric, and to delete the objectives that are not included within the most coherent plan. Notice that the most coherent plan, on some coherence metrics, may be one that adds new objectives to the list: the most coherent plan that unifies the already available objectives of surveying region A, and of surveying region C, may also involve surveying a region, B, between A and C, even though surveying B is not a means to either prior objective.[7] So cleaning up violations of representational hygiene can amount to doing (anyway some of) what we have already decided the Earth Rover needs to do: update and recompute its goals.

[7] For some experimentation with this approach, see Millgram and Thagard, 1996.

UU IR Author Manuscript

2

Creature-construction arguments resemble the more familiar category of state-of-nature arguments: in both sorts of argument, a streamlined version of a familiar phenomenon is exhibited as solving a simplified version of a pressing problem. The two types of argument differ primarily in that creature-construction arguments do not require the robots or organisms being discussed to appreciate the merits of the solution themselves, whereas state-of-nature arguments turn on the agreement of the population in the state of nature that the solution is satisfactory. In both sorts of argument, the point of exhibiting the artificial solution is to elicit a distinctive response, a recognition that has more or less this form: oh, so *that's* why we do it this way, and *that's* a way to see familiar, ordinary so-and-sos! We've reached the point in our own creature-construction argument where I need to invite that moment of recognition.

We have described an administrative device used to manage the activities of agents that have to navigate fluid and hard-to-anticipate environments. The device is constructed by designating a field of representations that is subject to a set of hygiene constraints, and then by implementing mechanisms that do a sufficiently good job of ensuring ongoing conformity of the representational field with the hygiene regime. We can now introduce two technical philosophical terms: *persons* are those administrative devices. The *logic* for a person is the system of informational hygiene constraints governing the designated field of representations. As we saw, the design solution involves the delegation of a particular level of responsibility: persons are *responsible for* getting their field of representations to conform to their logic. When it's people, rather than robots, these assignments of responsibility have a social dimension; we demand conformity to and sanction deviations from the hygiene regime.[8] And this last claim about how responsibility is assigned in the design approach we're now discussing matches our own practice. If I believe that p, and I believe that $\sim p$, I am logically inconsistent: because I am violating the principle of noncontradiction, I am required (or anyhow, this is the standard view of the matter) to alter my beliefs and remove the contradiction. However, if I believe that p, and *you* believe that $\sim p$, we merely disagree; there is no inconsistency on anyone's part, and no requirement on either of our parts to revise our beliefs. Our present way

[8] Brandom, 1994, ch. 1, provides a philosophically useful caricature of the system of penalties and demands. I expect that this approach to personhood can account for its moral and political dimensions as well; however, I will not take on those further tasks here. For preliminary groundclearing, see Millgram, 2009c.

UU IR Author Manuscript

of doing things enforces the logical norms only within the boundaries of a person (or an institution, such as a corporation, that is modeled on a person).

If we are right to think of a person as this sort of administrative device, we have to think of logic as an administrative technique as well. Thus, what persons are and what the laws of logic are turn out to be two problems that have to be solved jointly. Now, each problem on its own has given rise to a great deal of the wrong kind of philosophy: I mean, philosophizing that proceeds by table-thumping and insisting on whatever opinion one happens to already have.[9] But the coexplication of personhood and of logic—the idea that types of persons and corresponding logics come as packages—promises to give us much more traction on each of them: given what the field of representations is going to have to look like, it is reasonable to impose certain hygienic constraints, and not others; given what constraints need to be imposed, the field of representations must be demarcated in such and such a manner.

Two points are best registered before proceeding. First, the view of logic I am advancing is avowedly psychologistic, by which I mean that logic is understood to address the question of what the prescriptive rules of inference and reasoning are, and to do that by considering, among other things, what the mental activity of reasoning is like. For most of the twentieth century, accusing a philosopher of psychologism was a bit like accusing someone of communism during America's McCarthy era; it has been a long time since anyone has given level-headed consideration to the merits of psychologism. Allow me to recommend reviewing the arguments against psychologism, in both the analytic tradition (Frege) and the continental tradition (Husserl). You will find that they are fallacy-ridden, and that there were never any good (or even halfway-decent) arguments against psychologistic philosophy of logic.[10] I mean the present discussion to invite reconsideration of this long-neglected approach to logic.

Second, if persons are the administrative devices I have just described, persons are a design solution that will normally be only approximately implemented. Over the coming two sections, I will flag some of the ways in which actual persons deviate from the streamlined and idealized functional characterization I have given. For now, bear in mind that no one is likely

[9] Millgram, 2009b, sec. 7, briefly describes a representative example from mid-twentieth-century philosophy of logic.

[10] Kusch, 1995, and Ringer, 1990, pp. 295–298, provide historical and sociological background; Millgram, 2009a, pp. 223–225, discusses one of the arguments against psychologism.

UU IR Author Manuscript

to be more than an approximation to personhood. Persons are idealizations that are enormously important both for understanding what we are and for guiding what we do, but which we should not expect to find in nature.

Are we persons, even approximately? (I.e., is this the right idealization?) We have to get around in a world that is, by turns, familiar and deeply novel; even once we become familiar with some region or aspect of our environment, we are often forced to leave it and find our feet in a very different region of the world. We cannot get by on the goals we grow up with, or find in our animal repertoire; we have to map out our world, bit by bit, and change our plans as we see more of the map; when our plans change, our objectives change with them.

Naval vessels are described using lead ships that are of the same approximate design, as when we say that *Fletcher*-class destroyers were used in the Pacific theater. Adopting this convention, if the terse description of our capabilities that I gave a moment ago was correct, then we are Earth Rover-class agents. Earth Rovers can be expected to operate as persons, and we recognize the administrative procedure in our own activities. I think it is safe to presume that we are, anyway approximately, persons.

3

I need to pause to speak to an objection that many philosophers are likely not just to have but to feel very strongly about. I have described the hygiene rules introduced at the last stage of our creature-construction argument as a robot's logic, and I have presented this as a way of understanding *our* logic: an informational hygiene regime that can be justified as part of a particular design solution. But, the objection will run, whatever this is, it isn't *logic*. The hygiene constraints I described were negotiable; the design team might have settled on other rules than they did for this robot, or on very different rules for a somewhat different robot; they might decide at some point to upgrade the robot's list of constraints. Whereas, the objection continues, logic is a matter of fact, not something you can *decide* upon, and it is the same for everyone. (There cannot be one logic for Earth Rovers, another for people, and yet another for who knows who else.) The science of logic (what follows will be a bit of a caricature, but not much of one) is the investigation of the consequence relation; it is the investigation of the structural features of a realm of abstract objects; logic is a branch of mathematics. Logic can be studied, described and understood, but not *chosen* or *altered*.

Doing justice to the complaint requires the most dramatic available il-

UU IR Author Manuscript

lustration. The principle of noncontradiction was introduced above as an example of a choice the design team *might* make. For all I have said, they might have instead determined that cleaning up contradictions was not likely to be the best way of getting the robot through the challenges posed by its environment, and simply left noncontradiction off the list of constraints the robot is responsible for enforcing on its field of representations. On the view I am developing, even the law of logic that says that, for any p, $\sim(p \wedge \sim p)$ is a practical choice, made on the basis of performance considerations.[11]

On the traditional way of thinking, the bottom line is that both of p and $\sim p$ cannot be true, and that is all there is to it. But let's return to our Earth Rover mapping its environment, which may be the Sahara, or the rain forests of New Zealand, or Manhattan. If you think about it, it is obvious that different kinds of environments call for different kinds of maps. If the robot is on the city streets, it should construct a street map, but if it makes its way down into the subway system, it should construct a subway map. Its designers cannot know in advance what sort of map will be called for, and so, if they can, they will also delegate to the robot the task of determining what sort of maps to construct. Let's call the improved version that does this the *Earth Rover Advanced*; we can reserve the name "Earth Rover" for the simpler version that comes with a fixed set of map conventions.

It should be clear that because different types of maps will differ formally, different hygiene rules will be appropriate for them. For instance, a street map represents the distances from one point to another, and so testing for consistency of those distances with one another is appropriate. A subway map only needs to be topologically correct, and the distances from one point on the map to another will be properly determined by layout, readability and typographical requirements.[12] So the Earth Rover Advanced will also have to compute its own hygiene rules. That is, from the point of view of the creature-construction argument we are developing, there is no answer to the question: is logic fixed and unalterable? Rather, the question is how much responsibility has been delegated to the robot, and how much control is being reserved by the design team for the operators. At the Earth Rover stage, the

[11] It may even be a choice that was, in our own case, made incorrectly. In imposing these sorts of constraints, feasibility is normally a relevant consideration; your robot should be able to live up to its logic, more or less. Ancient and medieval logicians were unaware that truth-functional consistency is a computationally intractable problem; it is not a demand we can live up to, and in real life the constraint is only enforced locally, and never globally. (For Cook's Theorem, see Garey and Johnson, 1979, sec. 2.6.)

[12] Some subway maps try to have it both ways; it's instructive to compare a current New York City subway map with the much more usable (and incidentally aesthetically superior) 1970s MTA map.

UU IR Author Manuscript

design team delegates responsibility for constructing and maintaining a field of representations, while reserving to the programmers the specification of the constraints the field is required to satisfy; at the subsequent stage, the design team delegates responsibility for adjusting the constraints governing the field to the robot as well.[13] As before, the motivation in the former case is that the robot's designers know enough about the challenges posed by the robot's environment to specify its logic, but not enough to anticipate the concrete results to be obtained by enforcing the logic; in the Advanced case, the motivation is that they do not know enough about those challenges even to specify the robot's logic ahead of time. And so the question for us, as philosophers of logic, is: when we consider ourselves from the design stance, how much delegation of control should we take there to be?

In *Principia Ethica* (1903/1960), G. E. Moore pointed out that questions like, "It's pleasant, but is it *good*?" are always intelligible, in that they might, in principle, be answered one way or the other: they are, as he put it, open questions. He concluded that being good could not be, say, being pleasant (and more generally, that 'good' could not name any natural property); if it were, the question would not be 'open'. Today, Moore's Open Question Argument is remembered by metaethicists as historically important—it kicked off twentieth-century metaethics—and also regarded as thoroughly confused. It *was* thoroughly confused, but Moore had stumbled on an important philosophical tool, one whose proper use he

[13] Will the robot's logic be *formal*? As MacFarlane, 2000, has documented, that question has been taken to mean at least three different things over its history. But we can say this much: Constraints that are maximally general, rather than subject-specific, are suitable for guiding their robot through environments that are expected to contain subject matter that is unfamiliar to the design team. And constraints that can be applied simply by checking the forms of the representations, and which do not require judgment calls, are much easier to monitor, and so much more straightforward to impose. We can expect the design team for an Earth Rover-class device to specify a formal logic for it, if they at all can.

Can they? The design team may also anticipate a *messy* environment, one in which its robot's inferences will need to be driven by forced matches: for instance, by descriptions that are not fully true, but rather true *enough*. (See Millgram, 2009a, chs. 4–6, for a step-by-step exposition of this train of thought.) In the vocabulary of our creature-construction argument, that is to say that the robot must count some configurations of its field of representations as hygiene violations when they are not strictly ruled out, but come *close enough* to being ruled out. It is quite implausible that such a 'close enough' can be specified formally: for one thing, it is normally content- and task-sensitive. So a still more advanced robot, designed for an environment anticipated to be messy, will have a logic that may have a formal component, but which will in large part not be formal. I won't explore this point further here, but we should think of an exclusively formal logic as one step of a creature-construction sequence.

UU IR Author Manuscript

never understood. If we are right, asking such questions can serve, both in metaethics and philosophy of logic, as a probe that registers where in the creature-construction sequence one can find *oneself*.

That we are able to ask, "It's pleasant, but is it *good*?" tells us that we are not well-modeled by a robot whose designers understood its environment thoroughly enough to hard-code an evaluation of any particular sensation—this being Moore's naive model for pleasure—into the silicon. Instead, the right design, for creatures that have to cope with environments like our own, involves delegating the evaluation of sensations (and more generally, features of the 'natural' world) to the agent.[14] And if we are *un*able to see questions like, "It's good, but does that give me a reason to do it?" as sensible, that tells us that the further step to a particular more-ambitious delegation of control and responsibility was not necessary or feasible for *our* as-if designers. If we have these responses to this pair of questions, we are in a position to see ourselves as a design solution tailored to an environment fluid enough for the designer to need to delegate the task of evaluating one sort of item or another to his creation, but sufficiently well-understood for the designer to specify what the ensuing evaluations are to mean for the robot's decision-making. It might be necessary, when a robot has to operate in a still more fluid environment than that one, for its designer to allow it to recompute what its reasons are, *given* a set of evaluations; if we can see that last question as open, we should proceed on the assumption that we are better modeled by the robot at *that* stage of the relevant creature construction. Let's call this use of such questions *Moore's Probe*.

Return now to philosophy of logic. Philosophers who find it unintelligible to ask, "It's a contradiction, but is that a reason to revise my beliefs?" may think that their intuitions about the principle of noncontradiction are revealing to them a matter of fact belonging to a mind-independent science of logic. What they are really doing is locating themselves in the creature-construction sequence: they are identifying themselves as Earth-Rover-class, as opposed to Earth-Rover-Advanced-class agents.[15] The responses elicited by Moore's Probe are valuable, but not as limning a consequence relation to be found in a realm of Platonic entities; instead, they help us identify our own capacities and limitations.

[14] Why the scare quotes? Just try to get a philosopher to explain to you what he means by "natural".

[15] Philosophers whose use of Moore's Probe returns this result sometimes get huffy at the suggestion that there is another alternative, and here is a typical instance: "no attempt will be made to argue with those who think it acceptable to contradict oneself" (Williamson, 1994, p. 189; compare p. 136).

4

If you look at a directory listing on your computer, you will notice toggles on each file or subdirectory for read, write and execute privileges. The Earth Rover follows an administrative procedure which enforces informational hygiene constraints on a field of representations. How will the design team set the privileges on the items in that field?

The Earth Rover is assuming responsibility for correcting violations of the hygiene regime. It cannot do that unless it can identify violations. And it cannot identify violations unless it can check whether the entries in the field conform to the constraints. So the designers will assign it read privileges to the field of representations it is responsible for maintaining; that's just part and parcel of the implementation of this administrative device.

That the robot *just knows*—via what is likely a primitive but in any case a low-cost, routine and highly reliable operation—what the entries in its field of representations are is an implementation requirement: one way or another, the design team has to make sure this is a feature of their device. The philosophers' labels for this requirement being met are 'first-person authority,' and, more recently, 'self-knowledge'; that you *just* know what you think, in that very special way that others do not, is treated as a remarkable feature of the mind, and usually taken to require a very special sort of explanation, in the style of old-school metaphysics. Accordingly, a philosophical subspecialization has emerged whose members compete to provide such explanations. Just for instance, according to some of them, you have what might as well be a camera inside your head that tells you what you think. Again for instance, according to others, your pronouncements about what you think are to be construed as analogous to cries of pain, and are not properly reports at all.[16] But we can now see that it scarcely matters

[16]For sample 'detectivist' views, see Armstrong and Malcolm, 1984, pp. 108–137, Nichols and Stich, 2003, ch. 4. The label is due to Finkelstein, 2003, which is itself one of several developments of expressivism. Unsurprisingly, we also find denials that there is any such feature of the mind to be explained; Carruthers, 2011, is an example.

Sydney Shoemaker has devoted a substantial part of his efforts over an academic career to accounting for first-person authority. (See especially Shoemaker, 1996, but also Shoemaker, 1963, ch. 6.) With a handful of exceptions, the arguments fall into two groups: those that show that when an agent falls short of self-knowledge, its rationality will be impaired, and those that show how a rational agent can exercise first-person authority even when it cannot do the job by directly retrieving items stored in memory. I have found these very helpful in thinking about the functionality of an Earth Rover.

From the point of view of the design analysis we are developing, arguments in the first group show why first-person authority is an implementation constraint on Earth Rover-class agents. And arguments in the second group show that first-person authority does

UU IR Author Manuscript

how the implementation requirement is satisfied, as long as it *is* satisfied. And unless we imagine our robot's design team to be obsessed with elegant engineering, we can expect it to be satisfied in a number of different ways: there should be no *one* such explanation for first-person authority. The appearance of an occasion for old-school metaphysics very often arises from confusing a *demand* with a *fact*; for philosophical purposes, all that matters here is to explain the demand.

That the Earth Rover has write privileges for its representational field is also an implementation requirement. The Earth Rover is assuming responsibility for correcting violations of the hygiene regime. It cannot do that unless it can remove violations. And that entails either deleting, overwriting, or adding representations. A largish proportion of the recent discussion of self-knowledge turns on the observation that you can know what you think by making up your mind about it.[17] This describes a side-effect of what is, once again, not a fact but a requirement: the Earth Rover has to be able to write out changes to its data; if it cannot revise entries in its field of representations, it cannot maintain its hygiene regime.

---

not need to be implemented across the board as hard-coded memory access operations.

However, that latter group of arguments do have to be taken with a grain of salt. Some of the proposed workarounds impose high computational overhead (because they depend on the agent running through overly involved trains of ratiocination). When the overhead is too high, these will be poor or even unworkable implementation choices.

[17]Among such 'constitutivist' views, McGeer, 1996, argues that you know what you think because, once you have decided and announced what you think, you can live up to your announcement, by ensuring that your subsequent behavior conforms to it. (See also McGeer, 2007, and, for a related view, Bilgrami, 2006.) Moran, 2001, argues that you know what you think because, whenever the question arises, you can reconsider your opinion about the topic at hand, on its merits. The accounts are very close indeed, but we can distinguish them this way: Moran is focused on upstream inputs to your decision, and McGeer on downstream followons. The question of philosophical interest is raised indirectly by Lawlor, 2003, which points out that they *do* sometimes come apart: you make up your mind, but don't follow through. So what brings it about that the phenomena of interest to McGeer and Moran so often travel together? How is it that once we have made up our minds what we think, we frequently follow through, and live up to what we have said we think? How is it that when we follow through, it is often because we have made up our minds? In my view, this is best addressed as an implementation issue, via the sort of creature-construction argument we are now developing.

Moran seems to take his 'transparency' view to compete with detectivism, but this is evidently an error. If you find out what you think about some question by reconsidering it, the bases for your reconsideration—without loss of generality, we can take these to be the premises of an argument—are either available to you in the very same way, or in some different way. There is only so much reconsideration you can afford on any occasion, and no amount of it will staunch the regress. So sometimes you must know what you think, not by way of exercising your write privileges, but by exercising your read privileges.

The requirement that the Earth Rover coordinate its representations as specified by the hygiene constraints will not normally be met if other parties are also able to modify representations within the field. (E.g., if the robot tries to eliminate the conflict between representations with the forms p, $\sim p$ by overwriting the p-representation with a q, that will not ensure conformity with the hygiene constraints if, at the same time, another party is likely to be busily adding representations to the field that perhaps conflict with the q-representation.[18]) If you look once again at the long directory listing on your computer, you will see that it allows the access toggles to be flipped different ways for different users: for instance, a file might marked as readable and writable by the system administrator, as entirely off-limits to another group of users, and as read-only for you. We can anticipate the design team will enable write privileges (and, if the robot is going to be operating in a hostile environment, read privileges) for the robot alone.[19]

Persons, evidently, are private, and because I will presently introduce a related but contrasting administrative device, let's call instances of the traditional design solution *private persons*. Because human beings who implement such a design solution will end up shaping their intellectual apparatus around its features, truisms like, 'Only you can make up your own mind,' and 'Only you can think your own thoughts' will end up conceptual truths, and will seem to invite superlative explanations. They do not. Logical privacy is a side effect of adopting efficient descriptive conventions against the background of the stably implemented administrative device of private persons: that is, against the background of a good deal of *de facto* privacy.

Third, recall that the Earth Rover will sometimes resolve violations of its hygiene regime by developing and executing a plan for revisiting a site about which it seems to possess contradictory information; more generally, sometimes its method for correcting violations involves adding new objectives to its task list and taking action. There is no point in having the robot write subroutines for itself unless it can run them; accordingly, it is also part and parcel of the design solution that the Earth Rover have execute privileges to its representational field. But the Earth Rover will not be able to complete the tasks it sets for itself if other parties are able to interfere with its activities by running their own programs on it, and so the design solution will also reserve execute privileges solely to the robot.

We can now turn to making a distinction and adding a needed qualifi-

[18] That said, there may be other workable approaches. Wikipedia is much more consistent than the rationale I have just given would lead you to expect.

[19] For a biological version of the point about restricted read and write privileges, see Sterelny, 2004.

UU IR Author Manuscript

cation. I introduced persons as zones of logical responsibility, and I pointed out that read-write-execute privileges are necessary if the administrative device is to be effective. However, it is important to acknowledge that we are not merely persons: we do engage in the administrative activities I have been describing, but this is not everything we do, nor everything we are. For instance, your muscles are not elements of your field of representations, but they are nonetheless part of you. And the Earth Rover's computational engine may process a great many representations on which it does not enforce the sort of hygiene rules I have been describing. Let's call the computational activities of the Earth Rover *cognition*, and the representations that belong to the field on which it enforces its hygiene constraints its *mental states*—collectively, its *mind*. Mental states are, in this usage, a subclass of cognitive states; we should not expect the Earth Rover to have across-the-board read-write-execute access to all of its cognitive states. After all, we have no reason to think that, from the point of view of the design team, it needs those across-the-board privileges.

Let's explore this distinction. First, consider whether we should think of perceptions as mental or merely cognitive states. If they *are* mental states, they might seem to be an objection to the account I have been developing; we cannot change what we see and feel in the way that we can change our mind about other things, which is to say that we do not seem to have write privileges to our sensations. Now, that observation has to be complicated a bit, because we are told that the computational states that implement our sensations are in fact undergoing constant revision. What matters for our purposes is that the revision is not transparent to us, we do not control it, and we are accordingly not held responsible for it—anyway in the way we are held responsible for logical inconsistency. We can model perceptions as subject to an asymmetrical informational hygiene rule: the Earth Rover is required to rectify mismatches between its perceptions and elements of the field of representations we have already described, but not by altering the perceptions to match other data in the field. As before, this is reflected in our own practice: expectation may determine perception, as the catchphrase has it, but we are not held *responsible* for adjusting our perceptions to our expectations, and we do not merit labels like 'inconsistent' and 'irrational' when we fail to do so—on the contrary.[20] I mentioned earlier that

[20]In fact, the situation is trickier than that. Although Dennett, 1991, is now considered somewhat dated, it is deserving of continued attention for developing a series of arguments whose conclusions, however, the author seems to have misinterpreted. These arguments establish that a perceptual experience consists of a constantly updated stream of representations, no element of which has the sort of administrative status shared by other elements

UU IR Author Manuscript

we are best thought of as approximations to the administrative idealization of personhood; now that we have noted that perceptions are only partially covered by the sort of hygiene regime I was describing, we can choose to classify perceptions as cognitive rather than mental states, or to describe the way we handle sensory inputs as one way in which we are messier than the clean administrative device I was using as a model.

Memories are also liable to strike philosophers as another potential counterexample: a class of mental states for which we have read privileges but not write privileges.[21] We are not supposed to change what we remember when it proves to be inconsistent with what we believe on other grounds, and as with perception, we can accommodate this in either of two ways. We can classify memories as cognitive states rather than full-fledged mental states, or we can register that the highly simplified, idealized model that we arrived at in our creature construction abstracts away from much that goes to make up actual human beings: by acknowledging, that is, that we are only approximately persons.

However, as with perception, the facts on the ground are more complicated. We regularly replace our memories with synopses, and those synopses with synopses of *them*... and there is evidently a creature-construction explanation for that fact. Tersely: recall the suggestion that memories are added to a robot initially to support debugging; when a violation of the informational hygiene regime is identified, it will often be cheaper to look at the backtrace to find the source of the error than to revisit the sites of earlier observations. (No doubt memories will have other uses as well.) Now, even if memory is inexpensive, searching it is not. To keep the ever-increasing bulk of records usable, it will have to be continually redigested into short, simpler and easier-to-navigate versions; and it may be reasonable to do so on the assumption that, roughly, the further back in the past the event is, the less it matters to be able to retrieve details. The upshot will be a robot, the Mark M, whose memories *are* being revised on a regular basis, but—and this is what allows us to classify them as cognitive rather than mental

of the central field of representations in a robot of the Earth Rover's design. That is, no item in the stream *counts*, for purposes of the hygiene regime, as the final, designated sensation which other representations are required to match. Accordingly, the requirements of fit between sensory inputs and the designated field of representations in minds such as our own are much harder to formulate, in that the sensory inputs are moving targets.

[21] Perception and memory are not the only such states. Let's suppose that a member of the Catholic church is supposed to believe what the Pope believes on matters of doctrine. The member will not be able to resolve hygiene violations by changing the Pope's mind, not, anyway, in the way that the member can change his own. Once again, we are only approximations to persons.

UU IR Author Manuscript

states—these revisions are not subsumed under the set of responsibilites for maintaining the robot's field of representations, as I described it above.[22]

5

Now that we know what persons are, I want to speak briefly to a recent debate over their borders at a time. The argument has to do with whether minds can be 'extended'—that is, with whether or not all of your thinking goes on inside your head. The alternative is that some of what is properly your own mental or cognitive activity might take place in intellectual prostheses such as calculators or datebooks, or even in nonartifactual parts of your environment.[23] I am unhappy with the state of the discussion, but I will do the very minimium in the way of straightening up needed to get to the next stage of our own argument.

First, in this debate, the participants treat the phrases 'extended mind' and 'extended cognition' as equivalent—although some of them lean heavily on the more scientific *sound* of 'cognition'.[24] Recall, however, that we have distinguished these terms: your mental states are subject to a hygiene regime which you are responsible for enforcing; cognitive states are representational states that may or may not be subject to those hygiene rules. Second, the participants take themselves to be arguing over a question of fact: whether there *are* (or might be) mental or cognitive states outside people's heads. Taking the existence (or possibility) of cognitive states outside the head to be a question of fact is a mistake, but not one that concerns us here.[25] However,

[22] Philosophers worry about personal identity—about what makes you the same person you were yesterday—and you may expect the addition of an archive to the Earth Rover to be an entry point into a theory of what it is for Earth Rovers to persist over time. I think that a creature-construction approach to diachronic personal identity is promising, but this is not the place to try it out.

[23] For a recent exposition and defense of the position, see Clark, 2010.

[24] In somewhat the same vein, you find the latinate 'intracranial' replacing the ordinary English 'in the head'.

[25] "Cognitive science" was used as a rallying point for a research program derived from a defunct philosophical theory of mind, Hilary Putnam's functionalism (1975). Most of the argument about extended cognition, which I'm trying to sidestep, arises from a deep problem with that view. The idea at the center of the research program is to treat minds as computational systems; the problem is that computation is a purely formal notion. So where one draws the line around the physical region that is being represented computationally is left entirely to the discretion of whoever is constructing the computational representation. It is a question of notation and convenience, not of fact. So when Clark and Chalmers, 1998, announced that cognition could be extended, they were not making a discovery, but stumbling upon the vacuousness of computational functionalism.

UU IR Author Manuscript

it should be clear at this point that what count as a person's *mental* states is a design choice: in our own case, it is something that, subject to feasibility constraints, we can *legislate*.

Nowadays, most people carry wirelessly networked computers with them, pretty much at all times. (They think of these computers as telephones.) They store a great deal of information on those devices, and they do not, at least yet, censor the flow of information between their brains and their phones. As things stand, if there is a representation of the form p 'in your head,' and a representation of the form $\sim p$ on your phone, you may be likely to miss an appointment, but you are not thereby inconsistent. We could change that, by requiring you to enforce the hygiene rules not only on what is now your mind, but on the information stored on your mobile telephone as well; we could mandate that a violation of the extended hygiene regime renders you criticizable as inconsistent and irrational. And if we did that, your phone's memory, or part of it, would *count* as part of the field of representations for which you are responsible: it *would* be part of your extended mind. However, I do not find it profitable or instructive to argue over whether to make this change, anyhow on its own.

(Briefly, for those to whom this is an unfamiliar observation: Any physical object can be represented as a Turing machine—in fact, can be represented as implementing *any* Turing machine—and here's a trivial example to give you the idea. Let the state of your body this minute be a simple Turing machine's initial state, and let the state of your body the subsequent minute be the machine's halt state. Let the surface of your body this minute be one of two symbols from the machine's alphabet, and let the surface of your body the next minute be the other symbol. Set the transition function of the machine to map the initial state and the first symbol onto the halt state and the second symbol. Presto: for two minutes, you're a Turing machine!)

Recent attempts to resist the thesis of extended cognition have appealed to the practice of cognitive science (e.g., Adams and Aizawa, 2008, Rupert, 2009). But you shouldn't take the possibility of extended cognition to be settled by appeal to what cognitive scientists happen to do these days, when they're figuring out human or animal cognition. The leading idea of the program, again, is that cognition is computation, but computers are *devices*, and computation is an artifactual category. Devices are items we can alter and improve; you can no more argue that extended cognition isn't really cognition, because it looks different than what cognitive scientists have been looking at so far, than you could have reasonably argued in 1975 that cars can't come with fuel injection, because 'automotive science' had to that point proceeded on the methodological assumption that cars come with carburetors.

UU IR Author Manuscript

6

Here, on the other hand, is a modification to the administrative device we have been discussing that it is now realistically possible to implement, and which I will try to convince you is of philosophical and practical interest.

In human beings, the read privileges to the contents of someone's brain are restricted to that someone, and there's a feasibility consideration that dictates doing it this way: we do not know how to support read privileges for third parties. But third parties (these days, primarily corporations such as Google and Facebook and Symantec) *can* access the contents of a networked portable computer, such as a mobile telephone, or of cloud storage, in real time. Not only can a third party search for violations of specified hygiene rules, it can correct them. We already do rely on this way of handling errors in trivial cases; when a driver is following turn-by-turn directions, and he misses a turn, the navigation device recalculates the route and gives corrected directions on its own. It is not hard to imagine progressively more ambitious forms of outsourced hygiene control. Starting very close to where we are now: when the device determines that there is too much traffic for the two of you to meet at the restaurant you had planned, the scheduling application negotiates a different restaurant with your lunch date's scheduling application, makes reservations, gives you the appropriate turn-by-turn directions, and both of you find yourselves getting out of your cars, not where you had originally agreed, but in time for lunch. However, if something goes wrong, although this sort of delegation may provide you with an *excuse* ('there was a glitch in the scheduling software'), today the responsibility for getting things right (say, for turning up at the right place and time) ultimately rests with you. If you do not actually turn up, you apologize, where that consists in acknowledging that you are at fault.

That could change. Consider the following reassignment of responsibility for informational hygiene. You (but I will qualify the use of that pronoun in just a moment) are responsible for enforcing informational hygiene requirements over a brain-based field of representations. You are responsible for enforcing the coordination of the brain-based field with a specified field of representations on your mobile telephone; that is, you are responsible for making sure their contents match. But the responsibility for enforcing informational hygiene requirements on the networked device is outsourced to third parties. (For instance, if what you think you are doing today and what the phone-based schedule says differ, you are thereby inconsistent, and required to make them match; but Google checks your on-line schedule for consistency.)

UU IR Author Manuscript

In this way of assigning responsibility, when the error was one the third party was supposed to correct, and you seem to apologize for not turning up on time, 'I'm sorry' no longer acknowledges that you are at fault; your excuse is no longer an attempt to mitigate blame. Rather, saying that you are sorry merely expresses polite sympathy, in something like the way that an administrator says that he is sorry when it turns out that accounting has not cleared your paycheck: he is acknowledging that it is unfortunate, but what accounting does or doesn't do is just not his responsibility.

Persons are administrative devices. Private persons are the traditional version of this device, in which a single central field of representations is demarcated by coextensive read, write and execute privileges; these support the allocation of responsibility for enforcing representational hygiene to a single locus. The alternative allocation of responsibility we have just described is no longer a private person (and because 'private persons' was our label for the only model of personhood we have had, the alternative no longer counts by traditional lights as a person at all). For reasons I will get to momentarily, let's call the alternative the *minimal person*.

When I was describing the alternative assignment of hygienic responsibilities, I promised to qualify my use of "you". Recall that people structure their collective intellectual apparatus around features of their lives that they come to take for granted: pronouns like 'I,' 'you,' 'he,' and 'she' have built into them the presupposition that a *person*—a traditional, *private* person—is being picked out. If minimal persons come to replace private persons, we will have to adjust our intellectual apparatus, and not just our conceptual and analytic truths, but our pronouns, to accommodate the new state of affairs.[26] In order to avoid awkward circumlocutions, and the even more awkward attempt to introduce terminology suited to arrangements that are not yet in place, I stuck with the current but not-fully-appropriate pronouns. You can imagine scare quotes around them as you make the necessary adjustments.

Minimal personhood requires restricting logical privacy, which in turn requires surrendering a good deal of old-fashioned privacy. Mechanical telepathy is an available off-the-shelf feature for mechanical computers, but not for brains. Consequently, the person, reconfigured as I have described, is minimal *in virtue of* being extended; minimality is not a matter of how large or small the fields of representations being adminstered are, but of

[26]The general point is nicely made by Rovane, 1990; she is contemplating a society in which Parfit-style fissioning—people splitting into two—is an ordinary and anticipated occurence.

UU IR Author Manuscript

whether there is an administrative monopoly.

Sometimes a one-person business becomes a large, publicly-owed firm; once upon a time, a city built on seven hills became a Mediterranean empire. It would be a mistake to think that the founder and CEO *is* (really!) the firm, in something like the way it was a mistake when the Roman Empire was taken to be really the city of Rome. And it would be roughly that mistake to observe the human being at the center of the minimal person, to observe that he still retains administrative responsibility for his brain-based representations, and to conclude that the human being *is* (really, still) the person. The human being has become a component, albeit an essential one, of a more complex administrative device, and that is reflected in his hygienic responsibilities: because the brain-based representations violate the hygiene regime when they do not match the computer-based representations, third parties can make up the minimal person's mind—can *come to believe*—on his behalf.

7

Minimal persons are to private persons as the libertarians' minimal state is to the swiss-army-knife state we inhabitants of the developed world have become used to living in. In the libertarian utopia, a pared back organizational core is retained in order to enforce the drastically pared back rules of the game, and to prevent armed invasion; other functions, even those that have come to be regarded as standard responsibilities of government, are outsourced or simply left to take care of themselves. A libertarian minimal state might outsource peripheral military functions to contractors, privatize the post office, or eliminate government food safety inspections in favor of private quality assurance. In the minimal person, the outsourced responsibilities differ, but the arrangement is structually similar: the brain and the body it is in play the role of an organizing anchor for an extended region of personal sovereignty, in which the exercise of responsibilities formerly assigned to persons is turned over to external agencies.[27]

How should we think about the suddenly possible future in which private persons have been supplanted by minimal persons? That is, since private

[27]The large-scale social organization composed of minimal persons is evidently a novel form of feudalism, one in which vassalage penetrates the boundaries of the person. Plato thought that the form of justice was the same in the *polis* and the individual, and he took that as a constraint on his own political theory. Recent political theory rarely endorses or satisfies this constraint.

UU IR Author Manuscript

persons are what we have meant so far in talking of persons, how should we think of a world without persons?

As we take up this question, we need to remain aware of the obstacles to addressing it intelligently. First, as recent history teaches us, it is hard to tell what the uptake of a technical innovation is going to look like in real life.[28] We should assume that the picture we have to work with is wrong in all of its details.[29] Second, it is hard for us persons to see the choice between private and minimal personhood as one that *we* can make, and so, as a choice at all.[30] And third, the track record suggests that the collective choice will not be made for the reasons we philosophers are likely to see as the right ones. Think of what we might call *Google panopticism*, the startling emergence of new norms of privacy.[31] In the new way of doing things, the

[28]For instance, in the mid-1960s, it was clear that networked computers were on the way; Abbate, 1999, recounts how the internet arose out of a series of mistaken expert judgments as to what users would do with the connectivity: email, amazingly in retrospect, was believed to be an unimportant application. The ARPAnet became the kernel of the internet as we now have it because its architecture was flexible enough to support functionality unintended by its designers.

[29]And we should not assume that the picture we are considering *is* the future, details to one side. One alternative possibility is that the emerging technologies will reinforce rather than undermine the private person. As it is, the responsibilities we assign to persons are only sporadically enforced; we honor them to a great extent in the breach, and when we do live up to them, it is often enough from habit, rather than because we will be called on it if we do not. But social networking sites make the record of one's commitments public, and may over the long term make it very hard to walk away from them. Perhaps we have never really been persons, or even close approximations to them, and perhaps we will finally be forced to become the traditional, private persons that we imagined we were. (For this last suggestion, I'm grateful to Aubrey Spivey; Vogler, 1998, suggests that, if this happens, we will not like it one bit.)

[30]Nussbaum, 2001, highlights the perplexities of a formally similar choice considered by Plato.

[31]The phrase is adapted from Michel Foucault's adaptation of Jeremy Bentham's name for a prison he planned but never built. In Bentham's panopticon, the interiors of all cells would be visible from a central observation post. Foucauldian panopticism is an institutional rather than architectural innovation: in former times, punishments were draconian, but laws were sparse; in many modern institutions, rules cover even small details of behavior, violations incur relatively minor punishments, and checkpoints are frequent. Foucault discusses prisons, insane asylums, hospitals and so on (Foucault, 1995, Foucault, 1988, Foucault, 1975), but for an academic audience, universities are probably the most helpful example. Students turn in homework assignments, are quizzed on the reading, graded on their classroom discussion and so on. All of those grades are entered into a spreadsheet, and eventually merged into a grade that is sent to the registrar and compiled into a transcript, used to determine the student's honors status, his eligibility for privileges and so on.

Foucault made a big deal of the oppressiveness of panopticism, and no doubt my stu-

details of individuals' lives are monitored and recorded with an exacting thoroughness formerly reserved for celebrities and spies—but the records are anonymized before use. Google panopticism was accepted by the public largely because it was the price of a free, high-performance search engine, of the minor convenience of being able to pay with a bank card, and of satisfying the yearning, appropriate to high school but not thereafter, to be seen to be a member of a popular clique.[32] If we become minimal persons, it will probably be for equally bad reasons.

Nonetheless, we can consider the merits of such a move. Here I will restrict myself to one type of payoff. Informational and representational hygiene regimes in Earth Rover-class agents must be selected in view of feasibility constraints; one does not solve a problem by delegating a task to a device that it is unable to perform. Human beings are quite limited in their native abilities, and these limitations no doubt go a good deal of the way towards explaining what those hygiene requirements have been. Kantians like to think that our intentions are allowable only if they can be certified by the so-called CI-procedure, which they think of as a test of practical consistency; the most obvious reason that we do not in fact impose the requirement is that in most cases we are unable to say what would happen if, as lay people put it, everybody did that.[33] Bayesian epistemologists like the idea of having 'credences,' that is, of assigning probabilities to propositions, rather than believing them outright; they want us to ensure that our probability assignments are consistent, and they want us to update them on the basis of Bayes' Theorem. The most obvious reason that we do not in fact force these demands on people is that no one could actually live up to them. Lewisians believe that we should understand our conversations to involve a shared but invisible scoreboard; the semantics of our utterances are shaped by the constantly changing score.[34] We do engage in this sort of conversa-

dents feel that way, but in fact the reach of the technique is limited by its administrative overhead. *Someone* has to update all those entries in the file; if anything is going to be done on the basis of that file, *someone* has to do it. And every institution has personnel shortages.

The remarkable innovation of Google panopticism is that monitoring and followup no longer require human intervention. Many very small decisions, each tailored to the details of an individual's track record, can be taken without spending the time of a human administrator.

[32] *Was* it accepted? (That is, did the public understand the bargain they were making?) That's suddenly become a politicized question, but I can vouch for the sophisticates—I mean, computer scientists, programmers and electrical engineers, who *did* understand the bargain—having gone along with it in just this way.

[33] For a more precise rendering of the CI-procedure, see Millgram, 2005, pp. 90f, 141f.

[34] Lewis, 1983; and just to convey the general idea, the scoreboard is supposed include

UU IR Author Manuscript

tional practice on occasion, but we are not prudent to do so when anything important is at stake. Because we cannot keep track of the scoreboards we cannot see, conversation partners cannot verify that they understand the course of the conversation the same way; when the back-and-forth is not just friendly patter, conversational scorekeeping is a discourse management method beloved of shysters.[35] The most obvious reason we do not in fact require Lewis-style conversational scorekeeping is that we just can't do it properly.

However, if third-party contractors can perform monitoring and correction that unassisted humans could not, we are in a position to consider imposing more ambitious hygiene regimes. I have mentioned that we are not very good at anticipating the uses that will be made of technological opportunities, and so I won't try to say what I think those will be.[36] Instead, by way of conveying a sense of what such changes might look like, let me gesture first at a recently imposed hygiene regime of this generic type that is familiar; after that, I will rehearse Bernard Williams's description of another much older shift in administrative requirements.

We could require that all information kept on your personal computer be associated with some node of a singly-linked graph, and also that each node in the graph bear a mnemonic label. That is just the organization of data in a directory tree; rather than leaving compliance up to users, it has long been enforced by their devices' operating systems. Very recently, much of that task has been assumed by applications and rendered invisible to users, who often do not know how their files are organized. Demands of the kind I am envisioning might well have this look and feel to their users, but need not be restricted to the superficial organization of information.

As a illustration of just how *deep* such changes to our collective hygiene regime can go, I am going to borrow Williams's account of how what we might as well call the logic of time underwent a dramatic revision between Herodotus and Thucydides.[37] Before the transition, one could make claims

a vagueness parameter; if I say something that's *pretty* vague, that parameter resets to a value in the 'pretty vague' range, which in turn determines how a subsequent utterance, by me or by my conversation partner, is assessed: the subsequent utterance comes out true if it's true when it's also understood to be pretty vague.

[35] Millgram, 2009a, sec. 7.7, explains why; for now, when your cognitive capacities are swamped by the demands of a task, predatory interlocutors are bound to take advantage of that fact, and of you if they succeed.

[36] You might expect that I'm looking forward to the day when Google will check our credences for Bayesian consistency, but that seems to me to be unlikely: Bayesian updating is also a computationally intractable problem; see Cooper, 1990.

[37] Williams, 2002, ch. 7. I'm not going to weigh in myself on whether Williams has the

UU IR Author Manuscript

about past events without committing oneself to there being a determinate amount of time that had elapsed between the past event and the present: when you said that once upon a time, Minos ruled the waves, you had not yet said anything that entailed that there existed a n such that, n years ago, Minos ruled the waves. (Yes, there was a time when he did, but no *particular* time.) Because the way of thinking is so alien to contemporary sensibilities, it is worth emphasizing that what Williams is exhuming is not the thought that you might not *know* how long ago a past event happened; *we* have *that* thought. It is, rather, a logic of time less demanding than our own.

By the time we reach Thucydides, we are on familiar conceptual ground. To say that there was no time in particular at which an allegedly past event happened meant that it had not really happened at all. "Once upon a time" came to mean that it happened, not in the past, but in a fairy tale, and Williams nicely recovers the sense of bewildered outrage on the part of an older generation unable to fathom why the youth were so dismissive of what were suddenly merely 'myths'.

Now notice that this change would not have had much in the way of a practical point (and probably would not even have been feasible) if not for devices like calendars, that is, devices that allow one to record and calculate with temporal location. (If that sounds too hifalutin: calendars allow you to keep track of and to count the days between events.) When such devices become routinely available (and I want to stress the 'routinely'), what we can think of as the countback requirement for claims about the past became a social option.

That last example was meant to remind you that these hygiene rules are our *logic*. No doubt there are costs to surrendering much of our privacy—both what ordinary people have meant by that term, and the logical privacy of interest to philosophers—but I will not consider them here. What we stand to gain is the opportunity to reconsider what the laws of logic are, to ratchet up their demands, and so—if those demands are well-chosen—to equip ourselves, or rather, our minimal successors, to cope with more challenging environments than our own.

References

Abbate, J., 1999. *Inventing the Internet*. MIT Press, Cambridge.

history right.

UU IR Author Manuscript

Adams, F. and Aizawa, K., 2008. *The Bounds of Cognition*. Wiley-Blackwell, Oxford.

Armstrong, D. M. and Malcolm, N., 1984. *Consciousness and Causality*. Basil Blackwell, Oxford.

Bilgrami, A., 2006. *Self-Knowledge and Resentment*. Harvard University Press, Cambridge.

Brandom, R., 1994. *Making It Explicit*. Harvard University Press, Cambridge, Mass.

Bratman, M., 2006. *Structures of Agency*. Oxford University Press, Oxford.

Carruthers, P., 2011. *The Opacity of Mind*. Oxford University Press, Oxford.

Clark, A., 2010. *Supersizing the Mind*. Oxford University Press, New York.

Clark, A. and Chalmers, D., 1998. The extended mind. *Analysis*, *58*, 10–23.

Cooper, G., 1990. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, *42*, 393–405.

Dennett, D., 1991. *Consciousness Explained*. Little, Brown and Company, Boston. Illustrated by Paul Weiner.

Finkelstein, D., 2003. *Expression and the Inner*. Harvard University Press, Cambridge.

Foucault, M., 1975. *The Birth of the Clinic*. Random House, New York. Translated by A. M. Sheridan Smith.

Foucault, M., 1988. *Madness and Civilization: A History of Insanity in the Age of Reason*. Random House, New York. Translated by Richard Howard.

Foucault, M., 1995. *Discipline and Punish: The Birth of the Prison*. Random House, New York. Translated by Alan Sheridan.

Garey, M. and Johnson, D., 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York.

Grice, P., 1975. Method in philosophical psychology (from the banal to the bizarre). *Proceedings and Addresses of the American Philosophical Association*, *48*, 23–53.

UU IR Author Manuscript

Holland, J., Holyoak, K., Nisbet, R., and Thagard, P., 1986. *Induction: Processes of Inference, Learning, and Discovery*. MIT Press, Cambridge, Mass.

Ismael, J., 2007. *The Situated Self*. Oxford University Press, Oxford.

Kusch, M., 1995. *Psychologism*. Routledge, New York.

Lawlor, K., 2003. Elusive reasons: A problem for first-person authority. *Philosophical Psychology*, *16(4)*, 549–564.

Lewis, D., 1983. Scorekeeping in a language game. In *Philosophical Papers, Volume I*, pages 233–249, Oxford University Press, Oxford.

MacFarlane, J., 2000. *What Does it Mean to Say that Logic is Formal?* PhD thesis, University of Pittsburgh.

McGeer, V., 1996. Is 'self-knowledge' an empirical problem? Renegotiating the space of philosophical explanation. *Journal of Philosophy*, *93(10)*, 483–515.

McGeer, V., 2007. The moral development of first-person authority. *European Journal of Philosophy*, *16(1)*, 81–108.

Millgram, E., 1997. *Practical Induction*. Harvard University Press, Cambridge, Mass.

Millgram, E., 2005. *Ethics Done Right: Practical Reasoning as a Foundation for Moral Theory*. Cambridge University Press, Cambridge.

Millgram, E., 2008. Specificationism. In Adler, J. and Rips, L., editors, *Reasoning*, pages 731–747, Cambridge University Press, Cambridge.

Millgram, E., 2009a. *Hard Truths*. Wiley-Blackwell, Oxford.

Millgram, E., 2009b. John Stuart Mill, determinism, and the problem of induction. *Australasian Journal of Philosophy*, *87(2)*, 181–197.

Millgram, E., 2009c. The persistence of moral skepticism and the limits of moral education. In Siegel, H., editor, *Oxford Handbook of Philosophy of Education*, pages 245–259, Oxford University Press, New York.

Millgram, E. and Thagard, P., 1996. Deliberative coherence. *Synthese*, *108(1)*, 63–88.

UU IR Author Manuscript

Moore, G. E., 1903/1960. *Principia Ethica*. Cambridge University Press, Cambridge.

Moran, R., 2001. *Authority and Estrangement*. Princeton University Press, Princeton.

Nichols, S. and Stich, S., 2003. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Clarendon Press, Oxford.

Nussbaum, M., 2001. The *Protagoras*: A science of practical reasoning. In Millgram, E., editor, *Varieties of Practical Reasoning*, MIT Press, Cambridge, Mass.

Putnam, H., 1975. Minds and machines. In *Mind, Language and Reality*, pages 362–385, Cambridge University Press, Cambridge.

Ringer, F., 1990. *The Decline of the German Mandarins*. Wesleyan University Press, Hanover.

Rovane, C., 1990. Branching self-consciousness. *Philosophical Review*, *99(3)*, 355–395.

Rupert, R., 2009. *Cognitive Systems and the Extended Mind*. Oxford University Press, Oxford.

Shoemaker, S., 1963. *Self-Knowledge and Self-Identity*. Cornell University Press, Ithaca, New York.

Shoemaker, S., 1996. *The First-Person Perspective and Other Essays*. Cambridge University Press, Cambridge.

Smith, M., 1987. The Humean theory of motivation. *Mind*, *96(381)*, 36–61.

Sterelny, K., 2004. Externalism, epistemic artefacts and the extended mind. In Schantz, R., editor, *The Externalist Challenge*, pages 239–254, de Gruyter, Berlin.

Vogler, C., 1998. Sex and talk. *Critical Inquiry*, *24*, 328–365.

Williams, B., 2002. *Truth and Truthfulness*. Princeton University Press, Princeton.

Williamson, T., 1994. *Vagueness*. Routledge, New York.

UU IR Author Manuscript

Winograd, T., 1972. *Understanding Natural Language*. Edinburgh University Press, Edinburgh.

UU IR Author Manuscript